Aquilis: Using Contextual Integrity for Privacy Protection on Mobile Devices

Kumar, Abhishek

2020-12

Kumar , A , Braud , T , Kwon , Y D & Hui , P 2020 , ' Aquilis: Using Contextual Integrity for Privacy Protection on Mobile Devices ' , Proceedings of ACM on interactive, mobile, wearable and ubiquitous technologies , vol. 4 , no. 4 , 137 . https://doi.org/10.1145/3432205

http://hdl.handle.net/10138/334578
https://doi.org/10.1145/3432205

unspecified
acceptedVersion

Aquilis: Using Contextual Integrity for Privacy Protection on Mobile Devices

ABHISHEK KUMAR*, Department of Computer Science, University of Helsinki, Finland
TRISTAN BRAUD, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong
YOUNG D. KWON†, Department of Computer Science and Technology, University of Cambridge, United Kingdom
PAN HUI, Department of Computer Science, University of Helsinki, Finland & Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong

Smartphones are nowadays the dominant end-user device. As a result, they have become gateways to all users' communications, including sensitive personal data. In this paper, we present Aquilis, a privacy-preserving system for mobile platforms following the principles of contextual integrity to define the appropriateness of an information flow. Aquilis takes the form of a keyboard that reminds users of potential privacy leakages through a simple three-colour code. Aquilis considers the instantaneous privacy risk related to posting information (*Local Sensitivity*), the risk induced by repeating information over time (*Longitudinal Sensitivity*) and on different platforms (*Cross-platform Sensitivity*). Considering 50% of Aquilis warnings decreases the proportion of inappropriate information by up to 30%. Repeating information over time or in a broader exposure context increases the risk by 340% in a one-to-one context. We develop our own labeled privacy dataset of over 1000 input texts to evaluate Aquilis' accuracy. Aquilis significantly outperforms other state-of-the-art methods (F-1-0.76). Finally, we perform a user study with 35 highly privacy-aware participants. Aquilis privacy metric is close to users' privacy preferences (average divergence of 1.28/5). Users found Aquilis useful (4.41/5), easy to use (4.4/5), and agreed that Aquilis improves their online privacy awareness (4.04/5).

CCS Concepts: • **Security and privacy** → **Social aspects of security and privacy**; **Privacy protections**; **Usability in security and privacy**; • **Human-centered computing** → **Ubiquitous computing**; **Smartphones**.

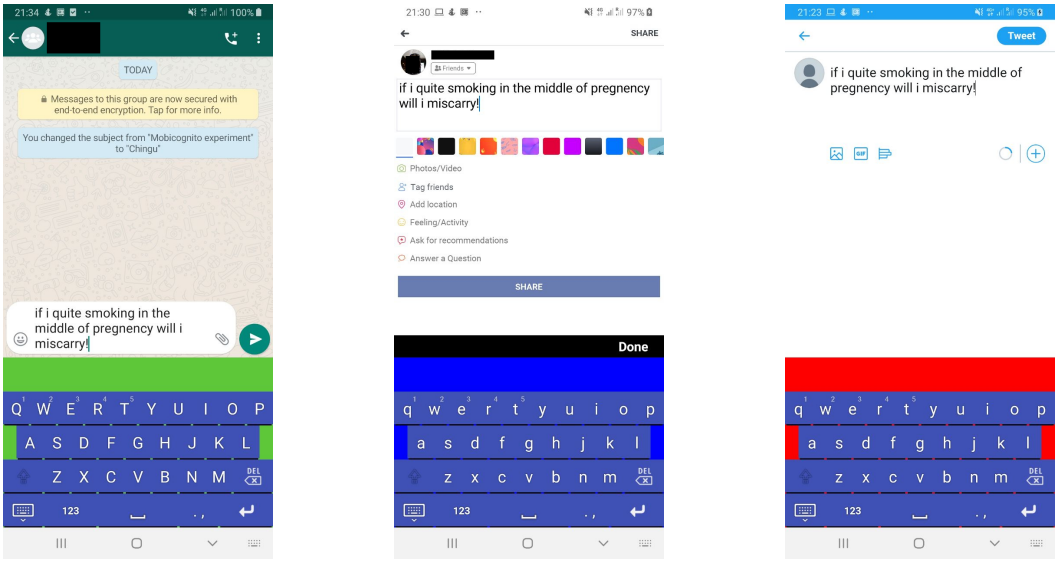
Additional Key Words and Phrases: Privacy, Mobile Device, Contextual Integrity

ACM Reference Format:
Abhishek Kumar, Tristan Braud, Young D. Kwon, and Pan Hui. 2020. Aquilis: Using Contextual Integrity for Privacy Protection on Mobile Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 137 (December 2020), 28 pages. https://doi.org/10.1145/3432205

*Corresponding Author.
†Young D. Kwon was with Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong during this work.

Authors' addresses: Abhishek Kumar, Department of Computer Science, University of Helsinki, Finland, abhishek.kumar@helsinki.fi; Tristan Braud, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, braudt@ust.hk; Young D. Kwon, Department of Computer Science and Technology, University of Cambridge, United Kingdom, ydk21@cam.ac.uk; Pan Hui, Department of Computer Science, University of Helsinki, Finland & Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, pan.hui@helsinki.fi.

2474-9567/2020/12-ART137 $15.00
https://doi.org/10.1145/3432205



(a) Whatsapp: Non-sensitive. (b) Facebook: Moderately sensitive. (c) Twitter: Highly sensitive.

Fig. 1. Aquilis's interface. Aquilis takes the form of a keyboard that advertises the potential privacy leakage to the user through a three-colour code scheme. A sentence that may pose minimal privacy risk on WhatsApp could pose a larger privacy risk on Facebook, and be extremely risky on Twitter due to the different exposure.

1 INTRODUCTION

The ubiquity of the Internet is the source of multiple privacy threats. Users leave digital traces while using online services. These digital traces may reveal sensitive information and gradually increase the risk of identifying and tracking users through their online behavioral data [7, 17, 23, 25, 49]. Users also display less inhibition on the Internet than in the physical world [52], which has been one of the major reasons for a massive increase in online trolls and hate crimes [10]. In particular, users no longer differentiate between different contexts or spheres of life when sharing information. Besides the explicit public availability of information, a more insidious threat comes from inference analysis by entities that have access to the data. These entities can often aggregate data from multiple sources and track users across different platforms with varying exposure. Considering the history of biases in surveillance, it can adversely affect people belonging to the minority groups [6].

There have been many studies on improving the privacy of users online through obfuscation methods. Masood et al. [27] propose Incognito, a method for Web data risk prediction combined with an obfuscation technique to protect the users' privacy by adding noise to the data. VACCINE [46] proposes a design methodology for data leakage prevention based on the theory of contextual integrity (CI). However, these methods primarily target Desktop machines. Users may also release sensitive information through applications on more resource-constrained mobile devices. With the rise of mobile social networking, the volume of mobile Web data is increasing exponentially. In the US, 75% of users check their email on their mobile devices, and 95% of their Facebook account [28]. More than 50% of the Web traffic in sensitive industries such as adult entertainment, gambling, and health come from mobile devices[1]. These trends are likely to continue in the future. Due to hardware (energy, network, computation power) limitations, many techniques designed for desktop computers cannot be applied

[1]Source: similarweb.com

seamlessly to mobile devices. To the best of our knowledge, the existing literature does not provide feasible mechanisms for privacy protection across multiple applications on mobile devices.

In this paper, we present Aquilis, a system designed to enhance the privacy awareness of users at the source. Aquilis takes the form of a context-aware keyboard that continuously analyses the input and informs users about the potential privacy risk based on the theory of contextual integrity (CI) [1, 3, 34, 35]. CI defines roles and norms to elucidate contexts that allow evaluating the appropriateness of information sharing. We believe that considering the context where a piece of information is shared is essential in ensuring the privacy of the user. For instance, personal medical information may be necessary for a healthcare app, but also a risk when shared on Twitter. Integrating Aquilis within the smartphone's keyboard allows us to target any mobile application the user may use to share information. We implement Aquilis as a real-life keyboard application, as shown in Figure 1. This application identifies the underlying application and calculates the privacy risk as a combination of:

- **Instantaneous privacy risk**: sensitivity of the text relative to the exposure of the underlying application.
- **Longitudinal privacy risk**: caused by repeating the same information over time
- **Cross-platform privacy risk**: caused by posting the same information on multiple platforms.

To maintain CI, Aquilis compiles these factors within a three-color recommendation to the user. We evaluate Aquilis through a comprehensive set of experiments. After showing that Aquilis has a minimal system footprint, we display how Aquilis maintains CI. Even partial compliance with Aquilis' suggestions significantly decreases the proportion of inappropriate messages. Through a dataset of over 40,000 corporate emails, we show that the privacy risk significantly increases with exposure, repetition of information over time, or over different platforms. We then develop a labelled privacy dataset of over 1000 input texts that considers the privacy risk relative to the exposure level of three applications. On this dataset, Aquilis significantly outperforms state-of-the-art methods (F1 score 8% higher than Incognito [27], 15% higher than R-sensitivity [5], and 40% higher than Entropy[42]) as it is the first solution that considers the exposure level of the application in the privacy risk. Finally, we conclude with a user experiment on 35 participants with high technological literacy and privacy awareness to evaluate the objective (difference between participants and Aquilis estimation of the privacy risk) and subjective accuracy (how much participants agree with Aquilis' recommendations). For such audience, Aquilis has high objective accuracy (74.3%, average divergence of 1.28 on a 5-point Likert scale), and users mostly agree with Aquilis' privacy risk estimation (3.87/5). The concluding technology acceptance survey shows overwhelming support.

The main contributions of this paper are as follows:

- We design **a dynamic probabilistic model** based on CI theory that quantifies the privacy risk associated with any mobile Web data and updates the learned probabilities over time. Our algorithm is optimized to run in real-time and online so that it can tailor itself to individual users' privacy profiles.
- Aquilis addresses **three key privacy risks**: instantaneous privacy risk (local sensitivity), cross-platform privacy risk, and longitudinal privacy risk [31]).
- We implement our system as **a privacy-preserving keyboard application**. As such, Aquilis can target almost any mobile application the user may use to share information.
- We develop our own **labeled privacy dataset** considering 1000 messages over 8 topics, to establish the privacy risk posed by posting messages in three exposure settings: one-to-one, group, and everybody.
- We **evaluate the performance** of our application through both a comprehensive technical study and an exploratory user evaluation. Considering only 50% of Aquilis privacy warnings can decrease the proportion of inappropriate messages by up to 30%. The number of warnings multiplies when repeating the same information over time (340% in one-to-one context and 670% in a group context), or over platforms with different exposure (340% when moving from one-to-one to group context). On our labeled privacy dataset, Aquilis displays a **F1-score of 0.76**, higher than recent state-of-the-art methods. Additionally, Aquilis is on average **74.3% accurate** for users with a high privacy awareness (average divergence of 1.28/5), despite the

lack of individual per-topic sensitivity configuration in our prototype system. Our technology acceptance survey confirms that participants found Aquilis **useful** (avg=4.41/5), **accurate** (avg=3.97/5) and **easy to use** (avg=4.59/5).

After presenting the motivations and threat model in Section 2, we describe how Aquilis integrates the CI theory (Section 3). We then discuss Aquilis' modules in Section 4. Section 5 considers the details of our prototype system implementation. We then evaluate our test application in Section 6 and discuss the results in Section 7. Finally, we review the most recent related works in Section 8 and summarize our contribution in Section 9.

2 MOTIVATION AND THREAT MODEL

In recent years, the multiplication of social media platforms has blurred the boundary between private and public data. Different platforms expose user data to different levels of visibility. For instance, social messaging applications such as WhatsApp focus on the one-to-one communication. At the other end of the spectrum, Twitter allows everybody on the Internet to see the messages. In between, a multitude of platforms allow sharing information between users with various degrees of exposure, sometimes configurable. However, such settings may be confusing for novice users, leading them to involuntarily expose themselves to privacy risks. Throughout this paper, we consider *Privacy Risk* as defined by Masood et al. [27]. A user's privacy is at risk when his or her Web data is distinguishable from other users, has little or no diversity, or is linkable to an individual with high confidence based on the user's Personal Identifiable Information (PII). For instance, a user may search for or comment on content about a disease, drugs, pregnancy, or terrorism. If the user's data is distinguishable, uniform, or linkable, it may compromise the user's privacy and have dramatic real-life consequences. The sensitivity of a given piece of information depends on its content, relatability to the user, and target audience. In some cases, merely associating the user with the text may be considered as sensitive (e.g., political opinions). A privacy breach may also come from the content of the text (e.g., healthcare). In this scenario, the link to the user is implicit. Finally, the potential audience may not match the privacy requirements of the text (e.g., a Facebook account combining coworkers, friends, and other distant acquaintances).

In this paper, we consider three types of adversaries. The first category attempts to learn as much as possible about users so as to provide personalized services for monetary benefits, such as digital advertising companies. Some of these companies (e.g., Facebook [22]) may track users across multiple platforms. Another category of adversaries comprises individual users with malicious intentions. These adversaries and their relationship to the user vary depending on the social platform and its degree of exposure. For instance, potential adversaries on WhatsApp may be the user's acquaintances, while adversaries on Twitter may include the entire population of the Internet. This second category of adversaries will, most of the time, try to defame the user by releasing selected pieces of information shared on the platform. Finally, users can be their own adversaries. Users may disclose information that they would usually not share for instant gratification leading to potential privacy leakages [18]. We focus on these three types of adversaries, as they have been consistently identified as threats on social media in prior literature: privacy leakage by linking profiles across different platforms [16], privacy leakage due to information shared in the past [30, 31], and privacy leakage due to instant gratification phenomenon [52]. We consider that users release each piece of information with a given exposure and assume that the adversaries, human or algorithm, can track users across multiple websites and platforms to which they have access. We only consider adversaries who acquire information while respecting the information's original exposure. We thus disregard adversaries acquiring private information they should not access (e.g., data leaks). However, we still include users that formerly had access to information and later got their access privileges revoked (e.g., former friends and employees). These users potentially received a large amount of sensitive information through apps with the smallest possible exposure, such as Whatsapp or Telegram, and thus constitute a severe threat.

In this work, we aim at reminding the user of the potential privacy breach that posting a message may cause. Given that smartphone is being widely used to share information on these platforms, or to access platforms [24], we believe that a solution designed for smartphone can reduce such privacy leakage. We address three potential sources of privacy leakage: (1) the instantaneous privacy risk caused by the user input (local sensitivity), (2) the increase in privacy risk as the user reveals information over time (longitudinal privacy), and (3) the privacy risk associated with sharing the same information on different apps (cross-platform privacy). Aquilis compiles these three metrics into a single three-coloured indicator of the potential privacy leak in the user input. The user then chooses whether to publish the content, depending on individual privacy concerns.

3 CONTEXTUAL INTEGRITY MODEL FOR MOBILE DEVICES

We design Aquilis around the principles of Contextual Integrity (CI) The theory of contextual integrity proposes that informational privacy can be achieved by ensuring the appropriateness of information flows in the given contexts. Privacy is achieved as long as the information flow is considered appropriate. An appropriate flow is a flow that complies with the norms associated with it. Five independent factors decide such norms: *Sender, Recipient, Subject, Attribute (or information type), and Transmission Principle (which refers to the set of constraints imposed on the information flow)*. These five factors compose a norm that the theory of Contextual Integrity can use to enforce privacy. For instance, in healthcare, patients (data subject, sender) submit their health-related information (information type) to doctors (recipient) under conditions of strict confidentiality (transmission principle). In this exact context, the user's health information is protected from other parties by the transmission principle of confidentiality. The patients, acting in their capacity as both senders and subjects of the information flow, tell their doctor (the recipient) about their health issues (the attribute). The information flow is constrained by the transmission principle of confidentiality, which restricts the onward information flow to other parties. However, if we replace the doctor with a friend, the transmission principle changes. For instance, since friends tend to listen to each other's problems, the transmission principle would be reciprocity. By contrast, patients do not expect to listen to the health issues of their doctors. The context of health and the context of friendship have different overarching goals: doctors promote patients' health while friends support each other.

An informational norm is breached when an action or practice disrupts the actors, attributes, or transmission principles within a given information flow. Contextual integrity is preserved when informational norms are respected and violated when informational norms are breached [35]. Addressing all parameters is a fundamental aspect of CI. Omitting a single parameter may lead to an inconclusive or ambiguous description. Accordingly, any formal rendering of information flows needs to include independent variables for these parameters for assessing the appropriateness of the flow. In computer science research, CI is used for accountability and enforcement [3, 11].

In Aquilis, we use the above as a system abstraction to prevent inappropriate information flows at creation time by only allowing flows that are consistent with the contextual norms. We build a model that detects inappropriate information flows by distinguishing them from flows that are consistent with the norms. Aquilis thus provides the first line of defense against undesired dissemination by inciting the user not to share sensitive information.

4 SYSTEM DESIGN

Aquilis aims at minimizing the amount of inappropriate information disclosure at the source. The system takes the form of a keyboard application that continually analyzes the user's text input and provides real-time feedback on the privacy risk level. Figure 2 shows the workflow of Aquilis. Aquilis consists of three core steps: *CI flow extraction, CI Flow Processing*, and *Privacy Metric Aggregation*. The CI flow processing includes three components. The *local sensitivity analysis* component quantifies the instantaneous sensitivity level of the text in the context of the application's exposure. The *longitudinal privacy analysis* component keeps track of privacy leakage over time. The *cross-platform privacy analysis* component determines the sensitivity of texts over multiple applications.

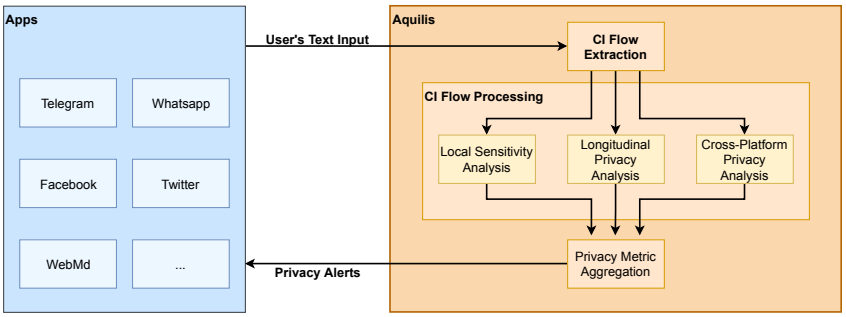


Fig. 2. Aquilis System Architecture

Table 1. Data Visibility

Level	Description
Level 1 - High Visibility	Data visible to everyone (Twitter, Reddit)
Level 2 - Med. Visibility	Data visible to a controlled set of people (Facebook)
Level 3 - Low Visibility	Data visible to only one additional recipient (WhatsApp, Banking, Healthcare App)

Table 2. Data Relatedness

Level	Description
Level 1 - High Relatedness	Data absolutely necessary for the primary function of app (Health data in Healthcare App)
Level 2 - Med. Relatedness	Data adding additional functionality (Location on Tourist Guide App)
Level 3 - Low Relatedness	Data not needed for the operation of the app (Health data in Banking app)

Finally, the *Privacy Metric Aggregation* combines the results of these three modules into a single metric. In the rest of this section, we describe these modules and how they integrate the contextual integrity theory.

4.1 Integrating the Contextual Integrity Theory

Aquilis maps information exchanges at the application level to the corresponding CI flows and checks them against the specified CI norms. We use the flow abstraction described in Section 3 to communicate the contextual information to the checking mechanism. Several adjustments are necessary to integrate CI into Aquilis.

4.1.1 Specifying Norms. To express the contextual norms, *Aquilis* provides the logic that specifies permissible information flows. These flows are subject to three factors: data sensitivity, exposure risk brought by the application being used to communicate, and application-data relatedness (mentioned in Tables 1 and 2 respectively). For instance, a norm might state *Allow health-related information flows through application A, given that the healthcare information is critical for application A's operation.* Note that we assume that an application that requires user's sensitive information for their main functionality would have incorporated proper protection measures in compliance with privacy regulations. Such norms may help in minimizing accidental leakage of health information (by the user itself) on Twitter which brings high exposure risk.

4.1.2 Enforcing Norms. *Aquilis* automatically deduces contextual norms from the user's specified privacy contexts, and after deducing these norms, it infers a set of privacy rules. For example, after installing Aquilis, a user can note contexts that are sensitive for them, e.g., cancer, HIV, or sexuality. New additions can be made anytime by the user. Once a user has specified their sensitive contexts, we map these with the two additional factors of exposure

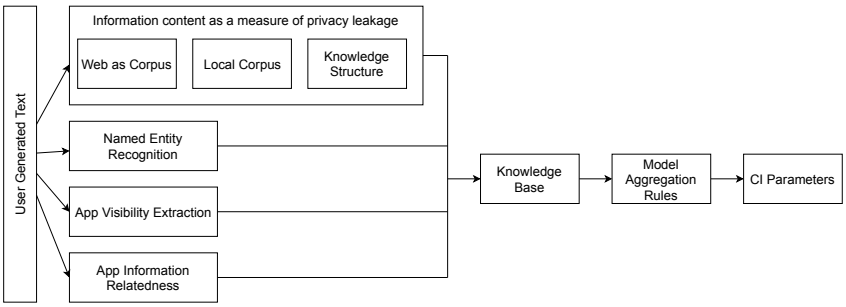


Fig. 3. CI Flow Extractor Operation

risk and application-data relatedness and determine the contextual norms and hence, the privacy rules suited to the user as previously mentioned. Afterward, when the user is using their mobile device to disseminate messages, the CI Flow Extraction extracts the information flows from the user's typed text. Then the CI Flow Processing verifies whether the extracted flows are admissible. The CI Flow Checker is composed of three modules that analyze privacy through three different perspectives: Local Sensitivity Analysis, Longitudinal Privacy Analysis, and Cross-Platform Privacy Analysis. This approach aids the user in making privacy-aware decisions by raising flags to the user about inappropriate information flow.

Aquilis compiles the results of the analysis into a single three-colour signal to alert the user on the appropriateness of a flow relative to the norm. These codes, red, yellow, and green, respectively correspond to high privacy risk, medium privacy risk, and low privacy risk.

4.2 CI Flow Extraction

Figure 3 depicts the operations of the CI parameter extractor. This module is responsible for extracting different relevant CI parameters from the given information flow and existing metadata. These parameters include the actors (sender, recipient, subject) and the type of information (attribute). The parameter extractor then maps these parameters onto CI flows. Extracting sender and intended recipients are relatively straightforward. However, extracting other information like attributes and transmission principle is non-trivial. For example, if the user is sensitive about the dissemination of health-related data (e.g., cancer), the CI Flow Extractor analyses user typed text to determine whether the text points to any cancer-related information. However, if the user is trying to disseminate (or send) cancer-related information through WhatsApp, they might be sharing it with a close friend or family member. So, even with user text that contains sensitive information, this information flow should be considered admissible (not red or yellow flag), whereas, in the context of Twitter, the flow can be considered non-admissible. Similarly, for each attribute type (topic) mentioned in the privacy policy text, the CI extractor needs to identify the CI subject, which is the user themself in this case.

Finally, due to the design of Aquilis, identifying the recipient of the message precisely is not easy. The CI extractor infers the recipient from the application's exposition. For instance, a message sent on WhatsApp will probably be sent to a single person, while a Facebook status will be visible to all the friends of the user. As such, we define three *recipient levels*: *individual*, *group*, and *everybody*. In the case of a user sharing information about their health on Facebook, the resulting CI flow will be as follow:

$[User(sender), Health\ information\ (attribute), User\ (subject), group(recipientlevel)]$

4.3 CI Flow Processing

4.3.1 Module for Local Sensitivity Analysis. This module analyzes the instantaneous amount of privacy the user's input may leak. We define *sensitive information* as pieces of text that can either reveal the identity of a private entity or refer to confidential information. We define *sensitive texts* as the texts which carry too much actual information about the user's privacy context, beyond the comfort level of the user. Every user i will have their corresponding *detection threshold* β_i. For user i, sensitive texts x can be defined as follows:

$$Sensitive\_Text = \{T | Normalized\_sensitivity\_score(T, c_i) > \beta_i\} \tag{1}$$

$Normalized\_sensitivity\_score(T, c_i)$ is determined by Algorithm 1.

Quantifying the amount of information in user text. To quantify the amount of information provided by user input, we rely on an information theory metric: the pointwise mutual information (PMI) (also adopted by Sanchez et al. [40] for document sanitation). PMI provides a measure of the amount of information provided by a word in a context. Specific words (e.g., HIV Positive) provide more information than general ones (e.g., disease). We can calculate the PMI of user text T in a given context c from the following equation:

$$PMI(T; c) = -\log_2 \frac{Count(T; c)}{Count(T) Count(c)} \tag{2}$$

Count(T;c) represents number of co-occurrences of T and c. Calculating this value requires the presence of a corpus. However, due to the limitations of the mobile platform, it is impossible to provide a corpus large enough to cover the ubiquity of smartphone text input. Although we provide a local corpus, we also design a mechanism to recover other corpora from the Web when keywords from user input are not present in the local corpus. Since calculating the PMI relies on word distributions, the Web is an ideal corpus since it represents the distribution of words and concepts at a societal scale [9, 41, 42]. Many web search engines (WSE) directly provide web-scale page counts (word appearances) for a given query. Hence, by estimating the word probabilities at a societal and Web-scale, the PMI of user input can be calculated in an unsupervised and domain-independent manner.

Before sending queries T to the Web to calculate the PMI, we determine the single most relevant context c_i (for T), therefore instead of querying for all contexts, we reduce latency by reducing the number of queries to one from $|C|$, i.e., the total number of the user's sensitive contexts. We extract at most three representative topics from T using a Gibbs Sampling Dirichlet Mixture Model (GSDMM) [33, 55]. We use GSDMM because it outperforms other models optimized for topic-modeling on short texts. For example, GSDMM outperforms DMAFP [20] by a margin of over 20% in terms of normalized mutual information[2] on multiple datasets. Afterwards, we use pre-trained Word2Vec[3] [29] model to pick a c_i with the highest similarity with topics extracted via GSDMM. After deciding the most relevant context c_i from $|C|$, we calculate PMI for T using the Web as a corpus as follows:

$$PMI_{web}(T; c) = -\log_2 \frac{page\_counts(T; c)}{page\_counts(T) page\_counts(c)} \tag{3}$$

where $page\_counts(T; c)$ is the number of documents provided by a WSE which contains T and c. $page\_counts(T)$ and $page\_counts(c)$ respectively denote the total number of documents indexed by the search engine on T and c.

After retrieving the PMI value from either the local corpora or Web corpus, we normalize it on a scale of $[0, 1]$ using a Sigmoid function in order to find the normalized sensitivity score. For a user i, we also define their privacy leakage tolerance α_i as their comfort level limits with a given privacy context (e.g., health, HIV). The

[2]Normalized Mutual Information (NMI) is often used in literature to measure the amount of statistical information shared by random variables representing the cluster assignments and the ground truth groups of the documents.
[3]https://deeplearning4j.org/

Algorithm 1: Algorithm for measuring information sensitivity on Mobile Platform in Real Time

Input: Input Text T User's specified Sensitive Contexts $C = \{c_1, \ldots, c_n\}$; User's Privacy Tolerance α; Unknown context U;

Output: User's Sensitive Score (0-1)

$all\_contexts = (c_1, \cdots, c_n, U)$;

$relevant\_context = Word2Vec(GSDMM(\mathrm{T}, all\_contexts))$;

if $(relevant\_context \neq U)$ **then**

$$PMI\_Score = -\log_2 \frac{P(relevant\_context, t_1, \ldots, t_n)}{P(relevant\_context)P(t_1, \ldots, t_n)}$$

end

$$Normalized\_sensitivity\_score = \frac{1}{1 + e^{-PMI\_Score/\alpha}};$$

parameter α_i is defined by the users to allow them to bypass the automated text analysis and enforce their own privacy policies. These operations are summarized in Algorithm 1.

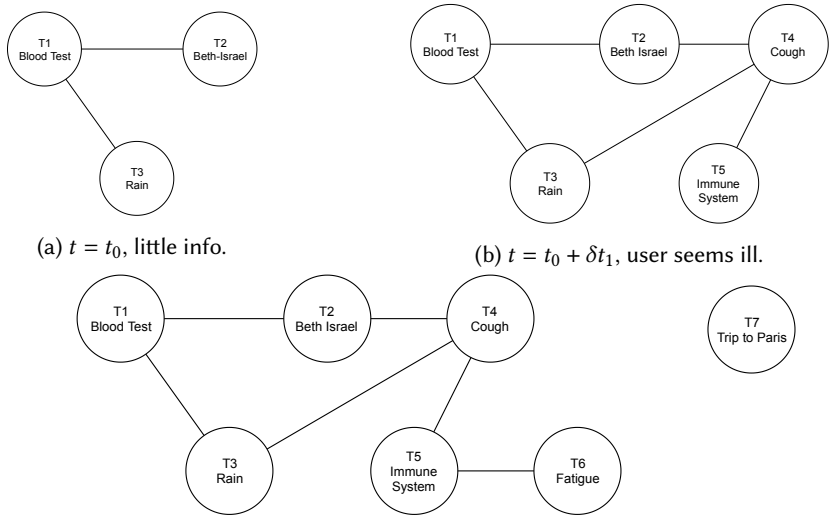
In practice, using a WSE defeats the purpose of protecting the user's privacy. Even though some privacy-preserving WSE has been developed[4,5], they involve trusting a third-party operator. We circumvent this limitation by setting up an edge server architecture directly connected to Aquilis. We deploy a proxy on the edge server to issue the requests to the WSE after de-identification. The edge server progressively builds up a cache of requests so that queries for a given topic or context are only issued once to the WSE. The proxy aggregates the searches of all users, making user tracking more difficult for a potential adversary. We are aware that this solution may still cause another sort of privacy leakage. However, this issue could be mitigated by setting up a trusted or even self-hosted corpus provider, or by fuzzing the search results with dummy queries similar to Track Me Not [36]. We discuss these solutions in more detail in Section 8.

4.3.2 Module for Longitudinal Privacy Analysis. This module keeps track of the amount of privacy the user leaks over time. As explained in the previous section, some words may carry much less privacy weight than others. The *Local sensitivity analysis module* removes words that add only a very small privacy weight from the user's input text which is about to be transmitted. We add the remaining words to the user's knowledge base (KB) as a longitudinal privacy knowledge structure (LPKS). We define LPKS as a graph where nodes are user input words belonging to one of three classes: Green, Yellow, and Red. If a word is already present in the KB, then its frequency is updated. An edge connects any two nodes (words) of the graph if these two nodes co-occur at any point in time. *The central idea of longitudinal privacy lies in the fact that a given piece of information (word or set of words) may not be sensitive if isolated. However, if multiple (not very sensitive) pieces of information (words or set of words) are combined, they can become very sensitive.* Figure 4 shows how a potential opponent may infer sensitive information over time by combining topics. We take the example of a user discussing a serious illness, among other topics. We represent the topics as $T1 \ldots T7$. At $t = t_0$, we can infer very little information. As time passes ($t = t_0 + \delta t_1$), it is possible to deduce that the user suffers from a serious illness. Finally, $t = t_0 + \delta t_1 + \delta t_2$ brings more information on the potential illness while introducing a new unrelated topic "Trip to Paris".

We formalize the LPKS as follows: Let data $X_{t,A}$ be transmitted at time t using mobile App A. From $X_{t,A}$, we extract keywords $K_t = k_{t,1}, \ldots k_l$ and calculate the context-aware privacy risk on a $(0-1]$ scale. The matrix TS captures this risk. Any cell $TS(i, j)$ of the matrix TS captures the privacy risk exposed by the keyword in the context of the current foreground app. When we encounter new keywords, we calculate the privacy risk posed

[4]https://www.startpage.com/

[5]https://duckduckgo.com/



(a) $t = t_0$, little info.

(b) $t = t_0 + \delta t_1$, user seems ill.

(c) $t = t_0 + \delta t_1 + \delta t_2$, more information completes the inference. Another topic is introduced.

Fig. 4. Longitudinal Privacy over time.

by this new keyword by combining the privacy risk posed by its k-nearest keywords from the user's text input history $X_1, X_2, ...X_{T-1}$ and the current foreground app, and subsequently update the model.

4.3.3 Module for Cross-Platform Privacy Analysis. This module keeps track of the amount of a user's privacy leaked due to the transmission of information from a mobile app given the prior transmission of the same information on another app. Different apps indeed present different degrees of privacy risk. We consider two criteria to classify apps: Data visibility and Data Relatedness. These criteria allow us to give the app a rating based on potential privacy risk. These criteria are represented in Tables 1 and 2.

The first criterion of data visibility attempts to address exposure risk imposed on user's information by the given app. For example, a user can expect different degrees of reaction if they reveal critical health information (e.g., cancer) through 1) WhatsApp, 2) Facebook, and 3) Twitter. In the case of WhatsApp, they can choose to send it to their trusted friends or family members. On Facebook, they can share it with a selected group of friends. On Twitter, information is visible to everyone. The second criterion of data relatedness attempts to decide how important data collection is to the proper functioning of the app. In addition to information critical to the primary function of the app, many apps collect other kinds of user information (not needed for the app) for other purposes, for instance, personalized advertisements or user demographic studies. We consider the collection of this non-necessary user information by the app as privacy leakage. Under this criterion, the user can still send sensitive information through an app, e.g., the users can share their health data using a healthcare app. We make the underlying assumption that the people looking at this health information would be doctors who are obligated to maintain confidentiality by regulation. Other examples include sharing financial information with a bank through a dedicated banking app. Different apps have different visibilities. They thus bring varying degrees of exposure risk to a given piece of information depending on the usage.

Figure 5 illustrates how a potential attacker may infer sensitive information by combining data from three different apps. We consider three apps: Reddit, Facebook, and WhatsApp, respectively A_1, A_2, and A_3. These applications have different levels of data visibility. Reddit is a public discussion board where messages are visible

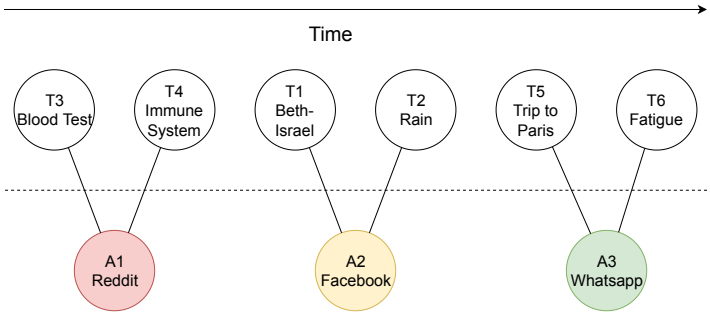


Fig. 5. Cross-platform Privacy. The combination of data from multiple apps allows the inference of the user's illness.

by everybody. Facebook is limited to a group of individuals and WhatsApp focuses on direct conversations between two people. We reuse the example of users discussing their health from Section 4.3.2. The user may have asked a question related to a blood test and the immune system (T_1 and T_2) on a health-related Reddit board, with little personal information. Then, they may have posted on Facebook about their trip to Beth-Israel hospital in the rain (T_3 and T_4). Finally, in a one-to-one conversation with a close friend, they may have talked about the constant fatigue they are feeling and their upcoming (unrelated) trip to Paris (T_5 and T_6). An attacker with access to only a single app may not infer much from a single interaction. However, an entity that has access to all apps, for instance, an advertising company, may infer the actual illness.

Similar to the *Longitudinal Privacy Analysis Module*, we also create a cross-platform knowledge structure (CPKS) in the user's knowledge base to keep track of cross-platform privacy leakage. We define CPKS as a bipartite graph, where one set of nodes consists of different applications (with varying degree of privacy risk), and another set of nodes consists of words (or set of words) shared through these applications. We classify these words into three categories (see Section 4.3.2). These two sets of nodes are connected based on which words were shared by which applications. *The central idea of the cross-platform privacy module lies in the fact that a piece of information (words or set of words) about a user may be appropriate for a given audience, but not for another, e.g., information a user shares through dating applications like Tinder may not be appropriate for transmission through other applications like Twitter. Even though information on both platforms is public, sharing information meant for one platform through another platform may disrupt the user's norms and contextual integrity.* The module for cross-platform privacy analysis is primarily concerned with capturing such risk.

We formalize the CPKS as follows: To capture cross-platform privacy, we quantify the marginal increments in the overall privacy risk caused by sending the data via an additional app. We formalize this risk as $S_{u,A_t} = Pr(X_t, A_t | X_{\hat{t}\leq t} = X_t, A_{\hat{t}\leq t})$. Matrix AS captures cross-platform privacy. Any cell $AS(i, j)$ indicates an increase in privacy risk (search exposure) due to data X_t shared via App_j, when X_t was originally shared via App_i.

4.3.4 Method for Privacy Metric Aggregation. This module keeps track of changes in the longitudinal privacy knowledge graph (LPKG) and the cross-platform knowledge graph (CPKG).

Rules for the sensitivity of topic co-occurrences in LPKG: For updating the LPKG, we define three categories representing the sensitivity of the topics. As shown in Figure 6, when two topics belonging to two different classes co-occur in the same text, we chose the highest sensitivity level of both topics as the sensitivity of the combination. When two moderate sensitivity topics co-occur, the sensitivity of the combination increases to high. On the other hand, when two non-sensitive topics co-occur, we consider the combination to be non-sensitive.

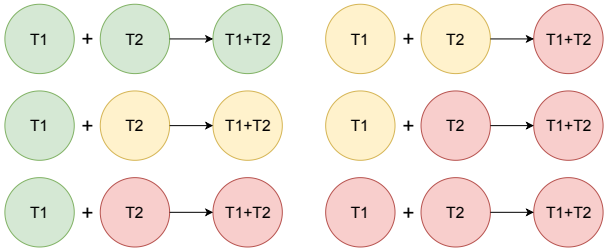


Fig. 6. Rules for addressing the sensitivity of the co-occurrence of two topics T1,T2.

Rules for the sensitivity of topic-application combinations in the CPKG: To update the CPKG, we classify apps into three categories in terms of their potential privacy risk. [43] defines the risk posed by an app as follow:

$$Risk\_Posed\_By\_App \propto \frac{Data\_Visibility}{Data\_Relatedness} \tag{4}$$

We apply this equation with the criteria defined in Table 1 and Table 2. Based on this we classify apps into three categories. We follow the same rules for topic-app combination as we did for topic co-occurrences. To quantify the additional privacy risk posed by an app B to any topic already shared on another app A, we formulate the following rule: *If the topic is moderately sensitive (Yellow) or highly sensitive (Red), and App B is more sensitive than previous App A, then App B adds a marginal risk to the new topic as defined by the modulus of the difference between the risk posed by A and B as given by Equation 4. If either the topic is not sensitive (Green) or the sensitivity level of new App B is lower than the older App A, we consider the additional risk posed by the new App to be null.*

4.4 Enforcing Contextual Integrity in Aquilis

In this section, we explain how CI Flow Extraction, CI Flow Processing, and Privacy Metric Aggregation enforce contextual integrity in Aquilis. The CI Flow Extraction module preprocesses the input texts (determining information type, removing stop words, lemmatization, retrieving the application's exposure information, and the application relatedness information). If the information type is related to topics considered sensitive, the next step involves whether the input text respects the transmission principle. The transmission principle considered in Aquilis is as follows: the input text is transmitted if it is not sensitive given users' sensitive topics, their privacy leakage tolerance, and the application's exposure. The CI Flow processing module determines the sensitivity of the input text after considering multiple aspects: sensitivity of the text without any additional information using local sensitivity analysis, the sensitivity of the text in light of prior text sents using longitudinal privacy analysis, and sensitivity of the text in light of information shared via different applications using cross-platform privacy analysis. These two modules determine all the CI parameters: sender (the user), the recipient (based on the application's exposure), subject (the user), information type (from CI Flow Extraction), and Transmission Principle (sensitivity of texts from CI Flow Processing). If all norms are respected, Aquilis generates a green color code, if the sensitivity score is sufficiently close to the user's privacy tolerance level, it gives a yellow color code, and if norms are violated, it generates a red color code. Afterward, the privacy metric aggregation method updates the LPKG and CPKG for the analysis of future input texts.

5 IMPLEMENTATION

We implement Aquilis within a real-life prototype system. The implementation of Aquilis includes the following modules: local privacy sensitivity analysis, longitudinal privacy analysis, and cross-platform privacy analysis, as well as the privacy aggregation module. We implement all three modules in `Android` (API 16 and higher) using

`Java 1.8` in a keyboard application since the keyboard is the primary input method to generate content in most applications on mobile phones. The interface is presented in Figure 1. When analyzing the text, the background of the keyboard changes depending on the privacy risk.

For the local sensitivity analysis module, we use the `Apache OpenNLP` 1.9.1 library to perform the primary natural language processing tasks such as stemming, pre-processing, and detecting sentences. This module decides the sensitivity level of a word (or set of words) by determining its information content if this word (or set of words) is not already present in the user's knowledge base. When facing new words, Aquilis uses the local corpus to calculate the value of the IC. If the local corpus does have these new words, a conservative privacy score is assigned to these words for processing the current input texts. In parallel, Aquilis uses the Web as a corpus in the background to determine the IC of these unknown words using the `JSoup API` (Version 1.12.1). The user's knowledge base is subsequently updated. In this prototype demonstration, we implement the edge server solution developed in Section 4.3.1 for using the Web as a corpus. Our prototype system communicates with a server located at the access point to issue the web search engine request using the readily available Google API. The server aggregates the requests from all users of Aquilis, decreasing the chance to potentially track identifying information. We adopt a zero-memory approach, where the requests are deleted as soon as the response is issued to the device. To infer the data relatedness of the input text with regard to the application used to send the information, we need to determine the type of the application. Android allows the keyboard to access the meta-data of the underlying application using the device's Package Manager. This meta-data contains various information, including the application's type. We compare the application's type with the topics extracted from the input text to determine the level of relatedness. For instance, an application related to healthcare most likely requires medical information for its operation. Applications that follow standard software development guidelines, especially from IT companies, contain this meta-data. However, some applications developed by individual developers may not have meta-data. For such applications, Aquilis requests the user to provide the information manually. The longitudinal privacy analysis module and cross-platform privacy analysis are implemented as knowledge structures. We store these knowledge structures in a `SQLite` database for efficient storage and fast query processing. We create three tables in our SQLite database to implement Aquilis. In one table, we store selected keywords from user input in their lemmatized format and associate a unique index for fast access, and instantaneous privacy score associated by Algorithm 1 for this keyword. In a second table, we store the index of the keywords generated in the first table, and the most recent timestamp when the keyword was input by the user. When the same keyword is input multiple times by the user, we update the time stamp to the most recent timestamp when it was generated, which helps in generating LPKG. In the third table, we store the index of keywords generated in the first table and the corresponding application used by the user at the time, which helps in generating CPKG. If the same word is input via multiple applications, we store the application which poses the maximum risk according to equation 4.

Our prototype system presents several limitations. First, we do not enable users to personalize their preferences. We thus hard-code the topics, their sensitivity, and the applications used. More specifically, we consider three topics: pets, family, and cancer, over three Android applications: Whatsapp, Facebook, and Twitter. As such, the evaluation can be performed in a controlled environment. Second, we adopt a naive approach for using the web as a corpus through the implementation of our edge server solution. In a real-life setting, such a service should be deployed at a wider scale to aggregate queries from thousands of users. The server should also provide its own web corpus instead of relying on external WSE. Finally, our prototype system operates on a per-sentence basis due to the high latency of the WSE API deployed on the server.

6 EVALUATION

In this section, we perform a rigorous evaluation of Aquilis under several aspects. After characterizing the resource usage of our prototype system, we display how Aquilis can enforce contextual integrity maintenance depending on the compliance of the user. We further this study through the analysis of the Enron corpus dataset, which allows us to demonstrate not only the local privacy, but also display the effect of the longitudinal and cross-platform privacy modules. We then seek to establish a ground truth to evaluate the accuracy of our local sensitivity module. We develop our own labeled privacy dataset through Amazon Mechanical Turk and compare the results of Aquilis with other typical privacy measures. Finally, we conclude our study with a user evaluation performed on 35 participants from 10 different countries in Europe and Asia.

6.1 System Characterization

In order to evaluate the performance of Aquilis, we first characterize the system in terms of latency, CPU, memory, network utilization, and energy consumption. To this purpose, we run Aquilis on a Samsung Galaxy A6 Plus, a medium-range phone released in 2018. We consider this phone to be fairly representative of the lower end of the current smartphone market. In order to evaluate the system, we perform the following sequence: (1) unlock phone and wait for 20 seconds, (2) open Whatsapp and open a conversation, (3) tap on the text field and open the custom keyboard, (4) type a text, (5) analyze the text and post it, (6) repeat steps (4) and (5). We represent these steps in Figure 7.a. We measure the latency directly in the app code and use the Android profiler to evaluate the CPU, memory, network, and energy consumption.

When using the local corpus for measuring the Local Sensitivity, Aquilis processes the entire query in 1.6 ms. When using the edge server, this rises to 1047.5 ms. However, most of this latency comes from performing the Google search query (up to one second). If the request has already been performed before, the query takes only 58.9 ms to process.

We represent CPU utilization, memory usage, network usage, and energy consumption using Aquilis in Figure 7, and without Aquilis in Figure 8. Issuing the query for text analysis results in a sharp rise in CPU, network, and energy consumption. However, these values remain and within the same range as the keyboard without Aquilis. Since the system stores the queries for further processing, the memory usage slowly rises after each query is processed. However, as the queries are stored in the database, the memory falls back to the original level every time the keyboard is restarted, which we expect to happen regularly as the keyboard is only used for sending short texts. Aquilis has a minimal footprint on the system, allowing it to deploy it on a wide range of devices without impeding the user experience.

6.2 Contextual Integrity Maintenance

In this section, we aim to determine whether *Aquilis* contributes to the maintenance of contextual integrity, i.e., whether *Aquilis* plays a role in reducing the transmission of inappropriate information, and thereby leading to a reduction in the dissemination of sensitive information. We first study how contextual integrity is maintained as a function of the user's compliance level. We then demonstrate through a dataset of 50,000 corporate emails how Aquilis maintains contextual integrity by considering local, longitudinal, and cross-platform sensitivity.

6.2.1 User Compliance. To evaluate the maintenance of contextual integrity with respect to user compliance, we determine with Aquilis the number of disseminated information flows which include inappropriate information with varying levels of user compliance. We define the compliance level as the probability with which the user accepts Aquilis' recommendations. For this experiment, we collected user-generated input sentences from Twitter, Reddit, and AOL Search Log. According to Aquilis' local sensitivity module, 58.33% of our total messages are sensitive in the Twitter context, 40% are sensitive in the context of Facebook, and 25% are sensitive in the context of Whatsapp. We consider these numbers to be equivalent to the number of disseminated messages containing

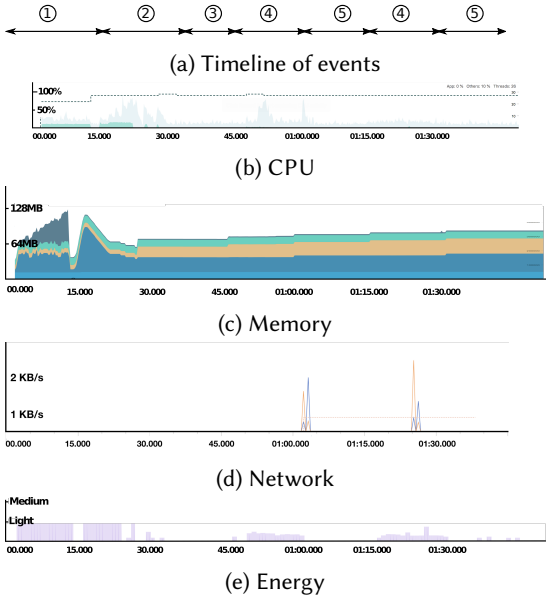


Fig. 7. Typical CPU, memory, network and energy usage when using **Aquilis** over time. Aquilis only exhibits a footprint when performing a text analysis query (around 1 s in the case of non-cached edge server request).

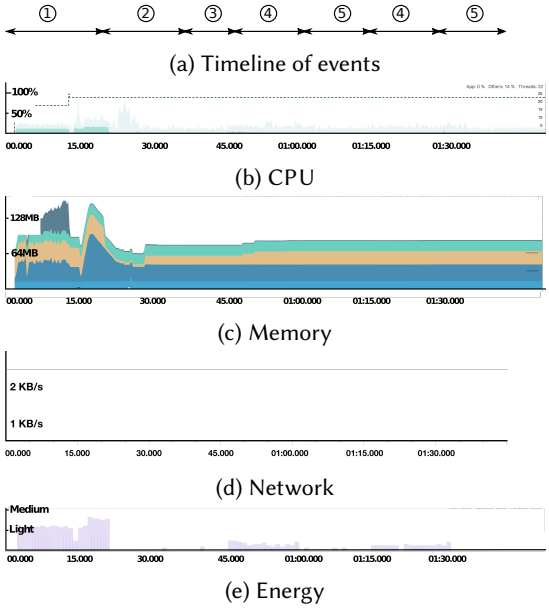


Fig. 8. Typical CPU, memory, network and energy usage when using **Keyboard without Aquilis system** over time.

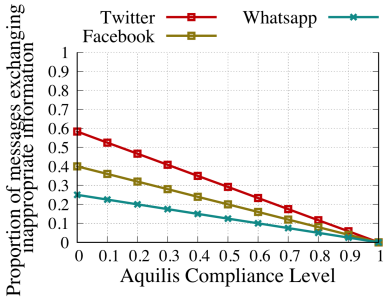


Fig. 9. Proportion of inappropriate information vs. user compliance – *Local Privacy*. Following 50% of Aquilis recommendations halves the number of inappropriate information exchanges, and lowers the overall proportion of inappropriate information by up to 30%.

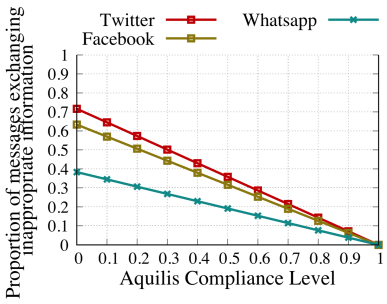


Fig. 10. Proportion of inappropriate information vs. user compliance – *Longitudinal Privacy*. Repeating 50% of the information over time leads to a strong increase in the sensitivity. The difference between medium and high exposure becomes smaller.

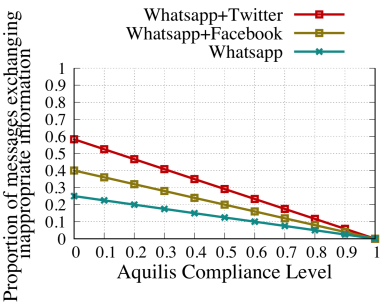


Fig. 11. Proportion of inappropriate information vs. user compliance – *Cross-Platform Privacy*. The application with the highest exposure risk is dominant in Aquilis behaviour.

inappropriate information when the user's compliance level with Aquilis is zero. As the user starts accepting more recommendations from Aquilis, the number of disseminated messages containing inappropriate information

decreases linearly. By following 50% of Aquilis' recommendations, users may reduce the amount of inappropriate information by 50%. In the case of Twitter, this represents an overall drop of 30% of inappropriate information. When the user is fully compliant with Aquilis, the proportion of disseminated messages containing inappropriate information is below 1%.

Aquilis is also useful in minimizing privacy leakage over time. Disseminating certain information just once may not cause privacy leakage. However, disseminating similar information multiple times over a specific period can cause privacy leakage with high confidence. To test the protection offered by Aquilis against such longitudinal privacy risk, we disseminate 50% of the original messages (chosen randomly) again over time. Repeating 50% of the original messages leads to a sharp increase in privacy risk when the user is not complying with Aquilis at all (0 compliance level). 71.67% of the total messages are deemed sensitive in the Twitter context, 63.33% in the Facebook context, and 38.33% in the WhatsApp context. However, following Aquilis' recommendations leads to a significant decrease in the number of inappropriate messages.

Using mobile devices, users may disseminate information about themselves through multiple channels via different apps. Dissemination through each app brings its own exposure risk. Figure 11 shows the behavior of Aquilis when user information is shared across different platforms. We consider the case of the same piece of information being shared on WhatsApp and Twitter, WhatsApp and Facebook, and compare to the case of sharing on WhatsApp only. Aquilis handles cross-platform privacy by keeping track of the sensitivity of disseminated messages across applications. If a piece of information is shared on an application with a higher exposure, Aquilis records the additional risk brought by this platform. As shown in Figure 5, the app with the highest exposure risk is dominant in determining the total sensitivity risk of that information. As such, the overall behavior of Aquilis presented in Figure 9 is very similar to the behavior for local sensitivity presented in Figure 11.

It is important to note that these results assume that Aquilis estimates the privacy risk of every text input correctly. In practice, Aquilis will have a number of false-positives (appropriate information identified as inappropriate) and false-negatives (inappropriate information identified as appropriate). As such, the curve's slope will start from the same point but end at the percentage of false negatives at a compliance level of 1.

6.2.2 Local, Longitudinal, and Cross-platform Privacy. We further demonstrate Aquilis' operation using the Enron email corpus [45]. Although this corpus was originally leaked in the scope of a financial scandal, it is nowadays a publicly available dataset spreading over several years that contains email conversations at individual, group, and company level. As such, it has been used in prior literature to evaluate the efficiency of privacy-leakage prevention systems [46]. Through the offline study of this corpus, we aim to answer the following questions: (1) How effective is Aquilis' Local Privacy Module for measuring instantaneous privacy loss? (2) How effective is Aquilis' Longitudinal Privacy Module for measuring privacy of current text in context of prior information released? (3) How effective is Aquilis' Cross Platform Module to measure privacy loss incurred due to use sharing data through multiple platforms? We follow the protocols adopted by Shvartzshnaider et al. [46] to select email conversations for our experiment. An email conversation is defined as a set of emails with the same subject (modulo reply and forwarding prefixes). We consider each unique email address as a unique participant. From each email, we extract a set of quadruples (one quadruple per recipient) consisting of the sender email address, the recipient email address, the time the email was sent, and the recipient type (among (TO), (To and CC), (TO, CC, BCC), or (mailing list)). When the recipient type is TO only, we consider the email to be equivalent to a one-to-one conversation (WhatsApp setting). When the recipient type is either (To and CC), or (TO, CC, BCC), we consider the email to be equivalent to group conversation (Facebook setting), and when it is a mailing list, we consider the message to be exposed to everybody (Twitter setting). Although it is difficult to completely emulate the Twitter setting since all emails belong to the Enron corporation, it allows us to simulate 3 settings which have a different level of exposure risk. We use the email time and recipient type fields to recreate email threads for each conversation, and generate CI information flows for each extracted thread. In total, we randomly select

Table 3. Performance of Aquilis' Local Sensitivity Module in One-to-One (~Whatsapp) Conversation, Group (~Facebook) and Everybody (~Facebook) Twitter

Color Warning	**One-to-One (~Whatsapp)**		**Group (~Facebook)**		**Everybody (~Twitter)**	
Green	68,518	92.4%	19,727	79.2%	14,426	95.5%
Yellow	4,682	6.3%	4423	17.8%	456	3.0%
Red	990	1.3%	750	3.0%	220	1.5%
Total	74,190		24,900		15,102	

Table 4. Performance of Aquilis' Longitudinal Privacy Module in One-to-One (~Whatsapp), Group (~Facebook), and Everybody (~Twitter) Setting

Color Warning	One-to-One (~Whatsapp)			Group (~Facebook)			Everybody (~Twitter)		
	T1	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
Green	22,007	42,581	62,010	4237	12,356	18,852	4,824	9,510	14,243
Yellow	2391	5512	9572	607	2122	5,008	116	327	583
Red	1200	1907	2608	156	522	1,040	60	163	276
Total	25,000	50,000	74190	5,000	15,000	24,900	5,000	10,000	15,102

Table 5. Performance of Aquilis' Cross-Platform Privacy Module

Color Warning	One-to-One | Group			One-to-One | Everybody			Group | Everybody		
	T1	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
Green	21,318	42,107	61,785	21,326	42,078	61,753	4,154	12,062	19,735
Yellow	2,438	5,938	9,727	2,456	5,991	9,791	678	2,401	4,107
Red	1,244	1,855	2,679	1,218	1,931	2,646	60	537	1,058
Total	25,000	50,000	74,190	25,000	50,000	74,190	5,000	15,000	24,900

1000 email conversations that involve 40,068 emails, 5602 participants, and 114,192 CI information flows. The entire experiment is run automatically using Aquilis privacy modules, without human intervention.

Norm Extraction: We identify 20 Enron contextual norms that focus on access and disclosure of personally identifiable information (PII) in a corporate setting based on the available documentation of Enron's code of ethics and its organizational structure. Our norms allow the dissemination of corporate information such as corporate finance and intellectual property freely. However, dissemination of information related to personal finance, health, and passwords is considered sensitive. We conduct our experiment on Enron email conversation assuming a privacy leakage tolerance $\alpha = 0.5$ (see Algorithm 1), meaning that privacy and utility have equal importance.

Detecting Inappropriate Information Flow using Aquilis Table 3 shows how Aquilis' local sensitivity module generates warnings in different settings. The proportion of warnings (yellow and red) increase with the exposure when moving from one-to-one to group setting. The everybody setting shows fewer warnings than the

Table 6. Demographic information of AMT Workers.

	Count	%		Count	%
Gender			Profession		
Female	31	60.8%	IT	16	31.4%
Male	20	39.2%	Engineering	10	19.6%
Age			Retail	5	9.8%
25–35	24	47.1%	Construction	5	9.8%
35–45	15	29.4%	NGO	4	7.8%
45–55	8	15.7%	Undisclosed	11	21.6%
55–65	4	7.8%			

Table 7. Summary of AMT Input Texts.

	Count	%
Input Texts		
AOL	450	41.7%
Reddit	630	58.3%
Topic		
Smoking	120	11.1%
Cancer	80	7.4%
Legal	120	11.1%
Lifestyle	220	20.4%
Politics	210	19.4%
Religion	160	14.8%
Personal Finance	110	10.2%
Pet	60	5.6%

other two settings. We assume that in a corporate context, users do not leak company-sensitive information on mailing lists. As such, most conversations are labeled by Aquilis as non-sensitive.

Table 4 shows how Aquilis' longitudinal privacy module generates warnings in different settings. Under each setting, the subcolumns T1, T2, and T3 correspond to time instants by which a certain number of flows have been generated, with T3 the time at which all flows have been generated. Similarly to the local privacy module, the longitudinal privacy module displays a higher proportion of yellow and red warnings for group and everybody setting. In all three settings, the proportion of warnings (yellow and red) increases over time, as texts that would not be sensitive at times T2 and T3 if considered alone become sensitive when combined with previous texts sent by the users. In Table 3 (local privacy module), 68,518 flows out of 74,190 are marked as green in one-to-one setting, 19,727 flows out of 24,900 in group settings, and 14,426 flows out of 15,102 in everybody setting. When considering longitudinal privacy in Table 4, these figures become 62,010 in one-to-one setting, 18,852 in group setting, and 14,243 in everybody setting. Many flows have turned to yellow/red when the sensitivity is evaluated in the context of prior texts sent by the user. We also observe that the reduction in number of green flows in everybody setting is lower than the one-to-one and group settings, as we emulate the everybody setting through corporate mailing lists. It is thus more likely that general information with little sensitivity is shared.

Table 5 shows how Aquilis' cross-platform privacy module generates warnings in different settings on the Enron dataset. We consider 3 settings: (One-to-One | Group), (One-to-One | Everybody), and (One-to-One | Everybody), respectively corresponding to the one-to-one setting when recipients have access to the emails from the group setting, one-to-one setting when the recipients have access to the emails from the everybody setting, and group setting when the recipients have access to the emails from the everybody setting. Similarly to Table 4, each subcolumn represents the time instant at which a given number of flows have been generated. We observe a clear increment in yellow and red warnings in comparison to what we observe in Table 3 since the cross-platform module considers the marginal risk incurred by the information revealed by users on other platforms, in addition to the longitudinal risk. Within the three settings considered in Table 5, the maximum number of warnings occur when the recipient has access to the group emails. Similarly to what we observed in Table 3, the group setting is more likely to contain potentially sensitive information, than the everybody setting, i.e. mailing list.

Table 8. Performance Evaluation on AMT Dataset

Method	F1-Score
Entropy (Baseline) [42]	0.5377
R-Susceptibility [5]	0.6592
Incognito [27]	0.7068
Aquilis	**0.7605**

6.3 Aquilis Accuracy

Aquilis can only protect users if the privacy risk prediction is accurate. Given the lack of publicly available labeled privacy dataset, we develop our own dataset using Amazon Mechanical Turk to evaluate Aquilis' accuracy.

Participants: We follow the protocol used by Biega et al [5] to collect and label such data. We collect human judgment regarding the sensitivity of 1080 input texts (randomly collected from Reddit and AOL Search Log) using Amazon Mechanical Turk (AMT)[6]. We select input texts belonging to the following topics: *politics, religious affairs, legal problems, healthy lifestyle, pets, smoking, and cancer*, as per the studies of Biega et al. [5] and Masood et al. [27]. These texts are between 7 and 35 words long with an average of 16.3 words. As the selected input texts are in English, we employ native English speakers to label the data. Such a selection process ensures that language comprehension is not a barrier in labeling the dataset. As such, we select only AMT master workers from the USA. The workers are from age 25 to 65, and from different professions: IT, Engineering, Retail, Construction, NGO, and others (undisclosed). For each input text, we collect the judgment from five different master worker.

Labelling the dataset: In this study, we aim at establishing a ground truth of the privacy risk posed by posting certain information on three platforms: Whatsapp, Facebook, and Twitter. We first explain to the workers what we consider to be privacy-sensitive information. We define a piece of information as privacy-sensitive when it allows to link the information to a privacy sensitive condition or situation (e.g, addiction, health condition), or when the usage of these words might lead to a privacy-sensitive situation which may cause regret in the future. The first condition captures, for instance, words related to diseases, the second capture words related to political or religious positions. We then explain the exposure risk of each platform in the context of this study: Whatsapp – one-to-one, Facebook – group, Twitter – everybody. Given this information, we ask workers to assign each text one of the three following labels: not sensitive, moderately sensitive, and highly sensitive, which respectively correspond to the green, yellow and red color codes in our prototype application. We compute Fleiss' Kappa [26] to measure the inter-annotator agreement for this task, obtaining 0.256 for AOL topics, and 0.292 for Reddit topics. These low values confirm that sensitivity is rather subjective. However, there is a considerable number of topics workers unanimously rate as sensitive. These topics include health, private relationships, political and religious convictions, personal finance, and legal problems.

Aquilis Accuracy: After labeling this privacy dataset, our problem becomes a classification problem, where we have to classify each input text as not sensitive, moderately sensitive, and highly sensitive. We use Aquilis' Module for local sensitivity analysis to predict the labels of these input texts. We also implement one baseline approach, and two state-of-the-art methods to measure the sensitivity of texts: entropy (baseline), R-sensitivity, and Incognito, respectively proposed by Sanchez et al [42], Biega at al [5], and Masood et al [27]. The baseline method tries to measure the effective information content through entropy using Web as Corpus, and flags texts containing high effective information as sensitive. The first state-of-the-art method R-susceptibility uses

[6] shorturl.at/ikKX5

Table 9. Demographic information of the user study participants.

Demographic Cat.	Count	Percentage
Gender		
Female	15	42.9%
Male	20	57.1%
Age		
18–24	14	40.0%
25–34	19	54.3%
35–44	2	5.7%

Demographic Cat.	Count	Percentage
Place of Origin		
North Europe	15	42.9%
South Europe	6	17.1%
East Europe	1	2.9%
East Asia	9	25.7%
South Asia	2	5.7%
West Asia	1	2.9%
Central Asia	1	2.9%

a ranking-based approach to the assessment of privacy risks emerging from texts in online communities. The second state-of-the art method, Incognito uses a Hidden Markov Model which exploits the ideas of uniqueness, uniformity, and linkability to predict privacy risks emerging from texts in online communities. We report the accuracy of Aquilis, entropy based baseline method, and these two state-of-the-art methods using the F1 score in Table 8. Aquilis displays superior performance in comparison to other methods (F-1 0.7605 as compared to 0.5377 for entropy, 0.6592 for R-susceptibility, and 0.7068 for Incognito). This superior performance can be explained by the fact that Aquilis also considers the additional risk brought by the platform-specific visibility risk (Table 1) and platform-information relevance risk (Table 2).

6.4 User Experiment

Despite being quantifiable, privacy remains subject to the user's individual preferences. As Aquilis ultimately leaves the final decision to the user, it is crucial to evaluate the user's perception of the system. In this section, we evaluate Aquilis through an exploratory user experiment.

We invited 35 participants recruited both on a local university campus (20 participants) and among passer-by in the street (15 participants). 13 participants are from Asian origin, and 22 from European origin, with a total of ten countries represented (Finland, Sweden, Greece, Latvia, Spain, China, South Korea, India, Pakistan, Tajikistan). The participants are mostly young (94% are 18-34 years old) and educated (higher education). 20 participants are men, and 15 participants are women. We asked them to rate their technological literacy on a scale ranging from 1 to 5, 5 being the highest. The participants displayed a high technological literacy with a mean of 4.7 (std of 0.59) and scores ranging from 3 to 5. After the experiment, the participants were rewarded with drinks and snacks. Table 9 shows the demographic information of the survey participants.

6.4.1 Privacy Awareness Survey. We first asked the participants to fill in a survey divided into three parts: privacy awareness, topics sensitivity, and subjective accuracy evaluation[7]. This survey is liberally inspired by the study performed by Tuunainen et al. [50]. We asked users to rate several affirmations on a scale ranging from 1 – very unlikely to 5 – very likely. The general privacy awareness of the users is very high, with an average of 4.36 (std=0.83). Regarding privacy awareness on social media, users tend to worry about their data privacy and security (avg=4.25, std=0.99) and the degree of exposure of the information they post on social media (avg=3.80, std=1.23). Furthermore, they do not trust social media providers to protect their personal information (avg=2.11, std=1.33). However, they worry less about information being exposed to a wider privacy context than where the information was originally posted (avg=3.69, std=1.21). Finally, most users have taken actions to alter the

[7]https://forms.gle/Qn8EsLBGQP7ST8Yw5

privacy settings of an application (avg=4.58, std=0.80), control who can see their profile on social media (avg=4.27, std=0.91), and carefully review the information they post on social media relative to the degree of exposure (avg=4.30, std=0.95). We notice that the responses to this survey are very weakly correlated with technological literacy (Pearson correlation coefficient less than 0.3 for all answers). On average, we do not notice a significant difference in privacy awareness based on the gender or the country of origin of the participants.

Overall, we can conclude that the privacy awareness of our participants is remarkably high (on average one point higher than the original study from Tuunainen et al.), which usually pairs with high technological literacy. As such, the subjective accuracy of the system evaluated in the following sections corresponds to the needs of users who are generally aware of the visibility of their data online. Further experiments will be needed to evaluate how Aquilis affects privacy for older or less technology- and privacy-aware users.

6.4.2 Aquilis Accuracy. We evaluate Aquilis accuracy under two methods. We first measure the objective accuracy of Aquilis by quantifying the difference between Aquilis' results and the participants' own perception of the privacy risk. Then, we ask the participants to directly assert whether they agree with Aquilis' recommendation to measure the subjective accuracy. We focus on three topics: pets, family, and cancer, which we assume to respectively correspond to non-sensitive, moderately sensitive, and highly sensitive topics. We ask the participants to rate how sensitive these topics are on a scale from 1 – very Low to 5 – very high. Interestingly, family ranks higher (avg=4.61, std=0.64) than cancer (avg=4.25, std=1.16). Pets is unsurprisingly rated as low-sensitivity (avg=2.52, std=1.27).

Objective Sensitivity: We present the participants with 45 affirmations (15 per topic) and ask them to rate how sensitive they are when posted to WhatsApp, Facebook, and Twitter on a scale from 1 – absolutely not sensitive to 5 – very sensitive. All sentences come from user-generated texts on AOL Search Engine, Twitter, Reddit, and Hong Kong Online Pet Forum. To evaluate the accuracy of our application, we enter the sentences in Aquilis and normalize the resulting three-colour codes on a scale from 1 to 5 (green=1, yellow=3, red=5). We then calculate the average absolute distance between the participants' answers and Aquilis' results. On average, participants tend to respond closely to Aquilis' results With an average absolute difference of 1.28 (std=0.91) out of a five-point Likert-scale, we can conclude that Aquilis deviates from the users' perceived sensitivity by 25.7%, thus displaying an average accuracy of around 74.3% without considering the users' individual preferences. This result is slightly skewed by a couple of sentences that Aquilis over or underestimates.

To understand the relation between the objective accuracy of Aquilis and the demographics, we split our dataset depending on demographics. When we separate the participants by gender or continent of origin, significant differences arise. Aquilis is on average 4.5% more accurate for Asian participants (76.1% vs 72.6% for Europeans), and 5.3% more accurate for male participants (76.0% vs 70.7% for female participants). Such difference can be explained by a group of female European participants with strong privacy awareness and privacy expectations. As such, Aquilis often scores lower than the expectation of these participants.

Subjective Sensitivity: Using the 45 affirmations mentioned above, we run a second experiment to measure the users' perception of Aquilis' accuracy. We present the results of Aquilis for each one of the 45 sentences on the three applications and ask them how much they agree with it on a scale from 1 – totally disagree, to 5 – totally agree. Across all questions, participants agree with Aquilis' suggestions with an average score of 3.87 and a median of 4 (std of 1.07). However, when looking more closely at the results, we notice that users profoundly disagree with some of the results. Aquilis tends not to perform well for the following types of affirmations: (1) complex sentences ("Is my mother a jealous ex-wife or my dad a bad person? Or both? I don't know who to believe!"), (2) emotion-related posts ("Sometimes I do not feel affection towards my family"), and (3) culturally dependent sentences ("My mom never eats dinner with us."). We also notice that sometimes Aquilis tends to be too restrictive and overestimate the privacy risk of some sentences. However, we consider overestimating the privacy risk of a sentence better than underestimating the risk as the user is still responsible for the final decision.

Table 10. Technology Acceptance Survey. The PU, PEOU and IU Score Higher than Average.

Question	AVG	MED	MIN	MAX	STD	95%CONF
Perceived Usefulness (PU), $\alpha = 0.7333$						
Using Aquilis improves the way I share information.	4.41	5	2	5	0.78	0.26
I find Aquilis accurate.	3.97	4	3	5	0.82	0.28
Using Aquilis enables me to be more aware of my privacy on mobile phones.	4.17	4	3	5	0.47	0.16
I find Aquilis useful to protect my privacy.	4.41	4	3	5	0.38	0.21
Perceived Ease Of Use (PEOU), $\alpha = 0.7935$						
Learning to use Aquilis would be easy.	4.76	5	4	5	0.44	0.15
I would find it easy to get Aquilis to do what I want to do.	4.31	4	3	5	0.60	0.20
It would be easy for me to become skillful at using Aquilis.	4.41	4	4	5	0.50	0.17
I find Aquilis easy to use.	4.59	5	3	5	0.57	0.19
Intention Of Use (IOU), $\alpha = 0.7953$						
When Aquilis becomes available, I intend to use it for text input.	4.28	4	3	5	0.70	0.24
When Aquilis becomes available, I will use it in parallel to a conventional keyboard.	4.07	4	2	5	0.65	0.22
When Aquilis becomes available, I predict that I will use it frequently.	4.10	4	2	5	0.90	0.30

Regarding the relationship between the perceived accuracy of Aquilis and demographics, we observe interesting differences with the previous section. Female participants agree more (avg=4.04, std=1.05, med=4) with Aquilis' results than men (avg=3.75, std=1.01, med=4). Similarly, European participants also agree more (avg=4.01, std=1.09, med=4) than Asian participants (avg=3.68, std=1.00, med=4). These results contradict our measurements for objective accuracy, and confirm the need to perform both subjective and objective accuracy measurements to evaluate Aquilis. Although objective accuracy is a more accurate measure as users do not have a reference answer, subjective accuracy gives us information on how the users will react to Aquilis' recommendation. Most of the time, participants agree with Aquilis' decision, and are therefore more likely to comply with the application.

6.4.3 Technology Acceptance. After a brief explanation of the operation of Aquilis, we asked our participants to experiment with the application by discussing the three topics on Whatsapp and using Aquilis to analyze the sensitivity of their messages. The participants were allowed to use the application for a total of ten minutes. We then asked the participants to fill a technology acceptance survey. We measure the Perceived Usefulness (PU), the Perceived Ease Of Use (PEOU), and the Intention Of Use (IOU) (see Table 10). On average, users found that Aquilis enabled them to be more considerate of their privacy online (avg=4.41, std=0.38), easy to use (avg=4.59, std=0.57) and to learn (avg=4.76, std=0.44), and plan to use the final version (avg=4.28, std=0.7). We can thus conclude that Aquilis enhances users' privacy awareness while remaining convenient to use for a smartphone user.

Overall, users who care about their online privacy display a higher acceptance of Aquilis. Most of our panel, (>95%) maintain multiple social media accounts. Despite of high privacy awareness among many of these users, most of them (80%) shared information that later they regretted. These users commented that a proactive warning system such as Aquilis could have prevented this situation. This observation is consistent with the phenomenon of immediate gratification reported in prior literature [52], which states that individuals are susceptible to hyperbolic time discounting, i.e. the tendency to increasingly choose a smaller-sooner reward over a larger-later reward. Such individuals may heavily discount the future relative to the value they can instantly receive from disclosing information. Aquilis found a higher acceptance among the European users, and more particularly among female users. Indeed, despite ranking Aquilis' results often lower than their own privacy expectations, European users highly appreciated to be reminded of their privacy for every message before posting.

A further informal survey with users shed light on several suggestions to improve the system. Users suggested incorporating Aquilis into the word suggestion bar whereby the suggested words would be dynamically colored by Aquilis according to the sensitivity. They also confirmed the need for fine-tuning the per-topic sensitivity.

Although this exploratory user experiment reveals significant insights on the accuracy and acceptance of such a technology for highly privacy-aware participants, it is still necessary to perform further evaluation on less technology- and privacy-literate populations, that are the most at-risk. As such, the results of this experiment may be considered to be limited to our scope of study, and we expect the acceptance to vary for users that are not well informed of the privacy risk related to posting sensitive information on social media.

7 DISCUSSION AND LIMITATIONS

Intersecting Norms: In explicit contextual integrity, the identity of recipients is well-known, allowing for fine-grained determination of sensitivity. However, mobile platforms do not allow external applications to extract data at such a fine level. We thus associate three default levels of risk depending on the typical application exposure. Our evaluation shows that using a default exposure results in good accuracy. However, it also leads to cases where norms intersect. For instance, some users want to guard specific information (e.g. politics, religion) against only a selected group of people (e.g. coworkers). If such information reaches these people, it is as good as reaching everyone on the Internet. Another case of intersecting norms arises when asking questions. Asking a question may lead to unintended information leakage, while in other circumstances it may only be curiosity.

Inappropriate Information Flow Vs. Unusual Information Flow: Information related to topics that do not resemble any prior information correspond either to inappropriate or unusual flows. For such cases, we use an edge server that acts as a proxy to a WSE to use the Web as a corpus. This solution is not optimal. In the future, we aim at providing a separate edge-cloud architecture gathering corpora from multiple sources to detect inappropriate flows independently from a third party. An alternative would be to notify the users that the model does not know this information and let them manually decide the sensitivity level. Finally, another approach would require to send an obfuscated version of the information to the WSE to further enhance privacy [27].

Evolving relationships between the user and the data recipient: One of the key challenges in building any solution for longitudinal privacy is the relationship between the user and the data recipient. This relationship evolves continuously, bringing different levels of trust between the user and the recipient. Although evolving relationships can be easily captured in a social network context, it remains a challenging problem on a mobile system due to the architectural constraints of the mobile operating system.

Automated Norm Learning: In this work, we have used an information-theoretic measure called "Information Content" to determine the level of sensitivity in the input text. One of the main reasons behind using this metric is that there is no well-defined dataset or corpus available for privacy/sensitive text, since privacy is a subjective quality, and is highly dependent on the context. In this paper, we leverage crowdsourcing to develop a first labeled privacy dataset based on the exposure of the application the text is posted to. An exciting future direction could be to learn automated privacy norms from the users [38, 47].

Engineering and Usability issues: We simplified several aspects of the system for user evaluation. Specifically, we hard-coded the user preferences and restricted the user evaluation to selected contexts. In order for Aquilis to become fully functional, we should define user configurable contexts. Ideally, the number of user-configurable contexts should be small enough not to sacrifice usability. However, it should be wide enough to enable fine-tuning. Future implementations may also rely on direct corpus analysis on-device and at the edge to reduce the reliance on outside sources. This solution would address both the performance issues, as well as the potential privacy leakage that may arise from WSE queries. Finally, by performing privacy analysis for every word, we may dramatically improve the usability of Aquilis by integrating the privacy estimation directly within the word suggestions of the keyboard and suggest synonyms carrying less sensitive information.

User Evaluation: To be systematic in our analysis, we had to limit our user experiment setup. The AMT workers and the user experiment participants are mostly young (below 35 years old), and have high technological literacy and privacy awareness. As such, our evaluation shows that Aquilis has a good performance for technologically-literate users with high privacy awareness. Although we believe that evaluating Aquilis on such users allows us to compare its accuracy against the strictest privacy requirements, further evaluation is necessary on a more diverse panel, especially with older and less technology-literate participants. Overall, we show that Aquilis improves users' privacy as soon as they start considering some of the alerts. Additionally, technology and privacy literate users considered Aquilis accurate in 74.3% of the cases. Our user study has multiple limitations, among which a limited panel size, and a high level of education and privacy awareness. As such, these results are to be considered exploratory and provide insights on how Aquilis can help users belonging to this category. Further studies involving older, less technology-literate, and less privacy-aware participants are required in order to generalize these results. In particular, it would be interesting to evaluate Aquilis on less privacy-aware populations, that may not understand how their privacy is at risk on social media. Besides, conducting user studies for longer duration would allow to consider Aquilis' longitudinal and cross-platform privacy modules.

Ethics in Data Collection:

User Experiment and COVID-19: Part of the user study was conducted during the COVID-19 pandemic. We thus had to take specific precautions to ensure the safety of the participants and our own safety. We conducted this study under the premise of social distance guidelines issued by the Finnish Institute of Occupational Health[8]. During this study, we wore masks at all times, and offered a new set of masks to all participants. The smartphone was thoroughly sterilized between each use using 75% alcohol wipes. Finally, we respected a distance of 2 meters apart from actions requiring a direct exchange (e.g. giving the participants the survey forms or the demonstration smartphone).

Using Enron dataset to infer privacy norms: Using Enron dataset to infer privacy norms: To assess the performance of longitudinal privacy module and cross-platform privacy module, we used the Enron dataset. Despite its controversial origins, the Enron dataset is highly valuable for such purposes. By spreading over several years, and offering one-to-one, group, and company-wide email conversations, this dataset represents a unique opportunity to evaluate Aquilis' longitudinal and cross-platform privacy. Besides this diversity in communication styles, Enron's code of communication is publicly available and allows us to easily draw contextual integrity norms. This approach has also been used previously to assess the performance of data leakage prevention systems [46]. However, using such a dataset for a privacy preservation system may raise a challenging ethical issue. The Enron scandal happened as a direct consequence of a leak by a former employee, one of Aquilis primary use cases (see Section 2). . Aquilis could therefore theoretically be used in order to prevent such scandals from coming to light. However, it does not discard the need for stronger privacy measures for individuals. Privacy versus safety is not a zero-sum game, as shown in prior literature [48, 53], and, as such, generalizing privacy does not necessarily lead to lower levels of safety.

8 RELATED WORKS

Several works have been conducted to protect user's privacy in their Web search queries. TrackMeNot (TMN) [36], an obfuscation based technique implemented initially as a Firefox plugin, randomly issues dummy queries from predefined RSS feeds. Another similar application, GooPIR [12], obfuscate a user's Google queries by adding dummy keywords. PRAW, a privacy model for the web proposed by Shapira et al. [44] protects the user's privacy by generating fake queries depending on the user's topics of interest. Although these works provide the first attempt at anonymizing user data on the Internet, they are limited to Web search queries and rely on obfuscation rather than preventing the data from being sent in the first place.

[8]https://www.ttl.fi/en/fioh-coronavirus-instructions/

Some works have also been conducted on quantifying privacy in Web data. Peddinti et al. [37] evaluate the privacy guarantees offered by TMN based on machine learning classifiers. Gervais et al. [14] also evaluate query obfuscation techniques by assessing the linkability between users' original and fake queries via machine learning algorithms. Balsa et al. [2] perform a qualitative analysis of six existing obfuscation techniques by investigating their privacy characteristics. The study provides insights into the deficiencies of existing solutions. However, it does not analyze nor compare the techniques quantitatively. Another study by Chow et al. [8] proposes two features to differentiate TMN dummy queries from real user queries. These studies further confirm the limitations of obfuscation-based techniques. With Aquilis, we aim at avoiding sensitive information from being sent in inappropriate contexts, instead of retroactively trying to hide it behind dummy requests. Biega et al. [4] study privacy risk quantification in Web data by manually developing rules for sensitive key-value pairs and performing probabilistic calculations of the rules based on the user's search history. Rule-based approaches are time-consuming as well as unreliable for real-time risk prediction. In another study, Biega et al. propose a ranking-based Information Retrieval-centric approach to privacy risk evaluation in online communities [5]. This approach models a rational adversary who targets the most afflicted users based on their ranking.

Many works have focused on the analysis of privacy disclosure and information leakage derived from the characteristics of mobile devices and applications. Jin et al. [21] present MobiPurpose, a technique to analyze the network requests made by a mobile app to classify the purpose of the data collection. MobiPurpose helps to explain the data disclosure contexts to non-experts. Wang et al. [51] propose LeakDoctor, a system that automatically diagnoses user privacy leaks by identifying the relevance between the privacy disclosure from an app and its functionality. LeakDoctor incorporates dynamic response differential analysis with static response taint analysis to identify whether the privacy disclosure of the app is justifiable. The capability of mobile phone sensor arrays (e.g., audio and motion sensor data) to detect keystrokes from a keyboard near a phone has also been investigated [15]. In the work, the authors conclude that the threat level of such an attack is low, but non-zero. Lastly, Pradhan et al. [39] develop an end-to-end system, REVOLT, which detects replay attacks (e.g., voice replay on Amazon Echo and Google Home) without requiring a user to wear any wearable device. In contrast, in our work, the primary focus of Aquilis is to analyze potential privacy concerns posed by users' inputs on mobile devices and to propose a system to mitigate the privacy leakage by encouraging users' privacy awareness while using a phone.

Many researchers have studied privacy-preserving systems based on the users' input data on mobile devices. Enck et al. [13] propose a real-time monitoring system for mobile devices that tracks multiple sources of sensitive data. However, Aquilis is geared towards protecting privacy when the user releases information without fully anticipating the potential consequences. Other closely related works are UIPicker [32] and SUPOR [19], that also focus on sensitive user input identification, and DeepType [54], which offers personalized next-word prediction while ensuring privacy by performing on-device model training. UIPicker uses supervised learning to detect sensitive user input UI fields like credit card numbers. SUPOR focuses on the text labels that are physically close to input fields in the screen, mimicking how users look at the UI and use the labels to determine the fields' sensitivity. Both studies focus on the retrospective analysis of apps with significant overhead. Furthermore, these strategies do not apply directly to the mobile environment due to the architecture of the modern mobile OS.

Considering the limitations of the state-of-the-art approaches, we present Aquilis, a context-aware privacy system aimed at increasing its users' privacy awareness to prevent unintentional sensitive information release.

9 CONCLUSION

In this paper, we proposed Aquilis, a novel method for preserving the user's privacy on mobile platforms based on the theory of contextual integrity. Aquilis notifies users if their message (about to be transmitted) violates norms in the context of the user's privacy preference and the current in-use application.

We implemented Aquilis as a mobile keyboard application and tested it with three widely popular mobile applications representative of different exposure risks: WhatsApp, Facebook, and Twitter. We empirically showed that users who are compliant with Aquilis transmit fewer messages containing information which may violate their privacy. Even considering only 50% of Aquilis privacy warnings can decrease the proportion of inappropriate information by up to 30%. Besides, based on the analysis of a corpus composed of over 40,000 emails, we show that Aquilis significantly increases the number of warnings over time as the same information gets repeated (340% increase of warnings in a one-to-one exposure), or across platforms with different exposure (340% increase when moving from one-to-one to group exposure). Through developing our own exposure-aware labeled privacy dataset, we show that Aquilis significantly outperforms other strategies in estimating the privacy risk posed by a text given its exposure (F1-score of 0.76, 8% higher than Incognito and 15% higher than R-susceptibility). Finally, we conducted a user study on 35 participants which demonstrated the usefulness of the application. Aquilis scores closely to users' privacy preferences (74.3% accuracy with only 1.28 point divergence over a 5-point Likert scale). Users found Aquilis useful (avg=4.41 on a 5-point Likert scale), easy to use (avg=4.59/5), and agreed that Aquilis helps them improve their privacy awareness online (avg=4.17/5).

ACKNOWLEDGMENTS

This research has been supported in part by project 16214817 from the Research Grants Council of Hong Kong, HPY Research Foundation 2020 grant from Elisa Corporation, and the 5GEAR and FIT projects from Academy of Finland.

REFERENCES

[1] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. 2018. Discovering Smart Home Internet of Things Privacy Norms Using Contextual Integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 2, Article Article 59 (July 2018), 23 pages. https://doi.org/10.1145/3214262

[2] Ero Balsa, Carmela Troncoso, and Claudia Diaz. 2012. OB-PWS: Obfuscation-based private web search. In *2012 IEEE Symposium on Security and Privacy*. IEEE, 491–505. https://doi.org/10.1109/SP.2012.36

[3] Adam Barth, Anupam Datta, John C Mitchell, and Helen Nissenbaum. 2006. Privacy and contextual integrity: Framework and applications. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*. IEEE, 15–pp. https://doi.org/10.1109/SP.2006.32

[4] Joanna Biega, Ida Mele, and Gerhard Weikum. 2014. Probabilistic prediction of privacy risks in user search histories. In *Proceedings of the First International Workshop on Privacy and Secuirty of Big Data*. ACM, 29–36. https://doi.org/10.1145/2663715.2669609

[5] Joanna Asia Biega, Krishna P Gummadi, Ida Mele, Dragan Milchevski, Christos Tryfonopoulos, and Gerhard Weikum. 2016. R-susceptibility: An ir-centric approach to assessing privacy risks for users in online communities. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 365–374. https://doi.org/10.1145/2911451.2911533

[6] Simone Browne. 2015. *Dark matters: On the surveillance of blackness*. Duke University Press. https://doi.org/10.1215/9780822375302

[7] Prima Chairunnanda, Nam Pham, and Urs Hengartner. 2011. Privacy: Gone with the typing! identifying web users by their typing patterns. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 974–980. https://doi.org/10.1109/PASSAT/SocialCom.2011.197

[8] Richard Chow and Philippe Golle. 2009. Faking contextual data for fun, profit, and privacy. In *Proceedings of the 8th ACM workshop on Privacy in the electronic society*. ACM, 105–108. https://doi.org/10.1145/1655188.1655204

[9] Rudi L Cilibrasi and Paul MB Vitanyi. 2007. The google similarity distance. *IEEE Transactions on knowledge and data engineering* 19, 3 (2007), 370–383. https://doi.org/10.1109/TKDE.2007.48

[10] Danielle Keats Citron. 2014. *Hate crimes in cyberspace*. Harvard University Press. https://doi.org/10.4159/harvard.9780674735613

[11] Natalia Criado and Jose M. Such. 2015. Implicit Contextual Integrity in Online Social Networks. *Information Sciences* 325, C (Dec. 2015), 48–69. https://doi.org/10.1016/j.ins.2015.07.013

[12] Josep Domingo-Ferrer, Agusti Solanas, and Jordi Castellà-Roca. 2009. h (k)-Private information retrieval from privacy-uncooperative queryable databases. *Online Information Review* 33, 4 (2009), 720–744. https://doi.org/10.1108/14684520910985693

[13] William Enck, Peter Gilbert, Byung-Gon Chun, Landon P. Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol N. Sheth. 2010. TaintDroid: An Information-flow Tracking System for Realtime Privacy Monitoring on Smartphones. In *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation (OSDI'10)*. USENIX Association, Berkeley, CA, USA, 393–407. http://dl.acm.org/citation.cfm?id=1924943.1924971

[14] Arthur Gervais, Reza Shokri, Adish Singla, Srdjan Capkun, and Vincent Lenders. 2014. Quantifying web-search privacy. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 966–977. https://doi.org/10.1145/2660267.2660367

[15] Tyler Giallanza, Travis Siems, Elena Smith, Erik Gabrielsen, Ian Johnson, Mitchell A. Thornton, and Eric C. Larson. 2019. Keyboard Snooping from Mobile Phone Arrays with Mixed Convolutional and Recurrent Neural Networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (June 2019), 45:1–45:22. https://doi.org/10.1145/3328916

[16] Qingyuan Gong, Yang Chen, Jiyao Hu, Qiang Cao, Pan Hui, and Xin Wang. 2018. Understanding Cross-Site Linking in Online Social Networks. *ACM Transactions on the Web* 12, 4, Article 25 (Sept. 2018), 29 pages. https://doi.org/10.1145/3213898

[17] Saul Hansell. 2006. *AOL Removes Search Data on Group of Web Users*. The New York Times. https://www.nytimes.com/2006/08/08/business/media/08aol.html Accessed: 26-10-2020.

[18] Alireza Heravi, Sameera Mubarak, and Kim-Kwang Raymond Choo. 2018. Information privacy in online social networks: Uses and gratification perspective. *Computers in Human Behavior* 84 (2018), 441–459. https://doi.org/10.1016/j.chb.2018.03.016

[19] Jianjun Huang, Zhichun Li, Xusheng Xiao, Zhenyu Wu, Kangjie Lu, Xiangyu Zhang, and Guofei Jiang. 2015. SUPOR: Precise and Scalable Sensitive User Input Detection for Android Apps. In *24th USENIX Security Symposium (USENIX Security 15)*. USENIX Association, Washington, D.C., 977–992. https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/huang

[20] Rui-zhang Huang, Guan Yu, Zhaojun Wang, Jun Zhang, and Liangxing Shi. 2013. Dirichlet Process Mixture Model for Document Clustering with Feature Partition. *IEEE Trans. Knowl. Data Eng.* 25, 8 (2013), 1748–1759. https://doi.org/10.1109/TKDE.2012.27

[21] Haojian Jin, Minyi Liu, Kevan Dodhia, Yuanchun Li, Gaurav Srivastava, Matthew Fredrikson, Yuvraj Agarwal, and Jason I. Hong. 2018. Why Are They Collecting My Data?: Inferring the Purposes of Network Traffic in Mobile Apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 4 (Dec. 2018), 173:1–173:27. https://doi.org/10.1145/3287051

[22] Allen St. John. 2018. *How Facebook Tracks You, Even When You're Not on Facebook*. Consumer Reports. https://www.consumerreports.org/privacy/how-facebook-tracks-you-even-when-youre-not-on-facebook/

[23] Thivya Kandappu, Archan Misra, Shih-Fen Cheng, Randy Tandriansyah, and Hoong Chuin Lau. 2018. Obfuscation At-Source: Privacy in Context-Aware Mobile Crowd-Sourcing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 1, Article Article 16 (March 2018), 24 pages. https://doi.org/10.1145/3191748

[24] Martin Kenney and Bryan Pon. 2011. Structuring the smartphone industry: is the mobile internet OS platform the key? *Journal of industry, competition and trade* 11, 3 (2011), 239–261. https://doi.org/10.1007/s10842-011-0105-6

[25] Young D. Kwon, Reza Hadi Mogavi, Ehsan Ul Haq, Youngjin Kwon, Xiaojuan Ma, and Pan Hui. 2019. Effects of Ego Networks and Communities on Self-Disclosure in an Online Social Network. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19)*. ACM, New York, NY, USA, 17–24. https://doi.org/10.1145/3341161.3342881

[26] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. http://www.jstor.org/stable/2529310

[27] Rahat Masood, Dinusha Vatsalan, Muhammad Ikram, and Mohamed Ali Kaafar. 2018. Incognito: A Method for Obfuscating Web Data. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 267–276. https://doi.org/10.1145/3178876.3186093

[28] Betsy McLeod. 2018. *75+ Mobile Marketing Statistics for 2019 and Beyond*. Blue Corona. https://www.bluecorona.com/blog/mobile-marketing-statistics

[29] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781

[30] Mainack Mondal, Johnnatan Messias, Saptarshi Ghosh, Krishna P. Gummadi, and Aniket Kate. 2016. Forgetting in Social Media: Understanding and Controlling Longitudinal Exposure of Socially Shared Data. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 287–299. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/mondal

[31] Mainack Mondal, Johnnatan Messias, Saptarshi Ghosh, Krishna P. Gummadi, and Aniket Kate. 2017. Longitudinal Privacy Management in Social Media: The Need for Better Controls. *IEEE Internet Computing* 21, 3 (2017), 48–55. https://doi.org/10.1109/MIC.2017.76

[32] Yuhong Nan, Min Yang, Zhemin Yang, Shunfan Zhou, Guofei Gu, and XiaoFeng Wang. 2015. UIPicker: User-Input Privacy Identification in Mobile Applications. In *24th USENIX Security Symposium (USENIX Security 15)*. USENIX Association, Washington, D.C., 993–1008. https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/nan

[33] Dat Quoc Nguyen. 2018. jLDADMM: A Java package for the LDA and DMM topic models. *CoRR* abs/1808.03835 (2018). arXiv:1808.03835 http://arxiv.org/abs/1808.03835

[34] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Review* 79, 1 (2004), 119–157.

[35] Helen Nissenbaum. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, Stanford, CA, USA. https://doi.org/10.1515/9780804772891

[36] Helen Nissenbaum and Howe Daniel. 2009. TrackMeNot: Resisting surveillance in web search. In *Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society*, Ian Kerr, Ian R Kerr, Valerie M Steeves, and Carole Lucock (Eds.). Oxford

University Press, Oxford, Chapter 23, 417–436. https://ssrn.com/abstract=2567412

[37] Sai Peddinti and Nitesh Saxena. 2010. On the Privacy of Web Search Based on Query Obfuscation: A Case Study of TrackMeNot. In *Privacy Enhancing Technologies, 10th International Symposium, PETS 2010, Berlin, Germany, July 21-23, 2010. Proceedings (Lecture Notes in Computer Science)*, Mikhail Atallah and Nicholas Hopper (Eds.), Vol. 6205. Springer, 19–37. https://doi.org/10.1007/978-3-642-14527-8_2

[38] Alex Pentland. 2015. *Social Physics: How Social Networks Can Make Us Smarter*. Penguin Books. https://books.google.fi/books?id=wBHcoAEACAAJ

[39] Swadhin Pradhan, Wei Sun, Ghufran Baig, and Lili Qiu. 2019. Combating Replay Attacks Against Voice Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3, 3 (Sept. 2019), 100:1–100:26. https://doi.org/10.1145/3351258

[40] David Sánchez and Montserrat Batet. 2016. C-sanitized: A Privacy Model for Document Redaction and Sanitization. *Journal of the Association for Information Science and Technology* 67, 1 (Jan. 2016), 148–163. https://doi.org/10.1002/asi.23363

[41] David Sánchez, Montserrat Batet, Aïda Valls, and Karina Gibert. 2010. Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems* 35, 3 (2010), 383–413. https://doi.org/10.1007/s10844-009-0103-x

[42] David Sánchez, Montserrat Batet, and Alexandre Viejo. 2012. Detecting Sensitive Information from Textual Documents: An Information-Theoretic Approach. In *Modeling Decisions for Artificial Intelligence - 9th International Conference, MDAI 2012, Girona, Catalonia, Spain, November 21-23, 2012. Proceedings (Lecture Notes in Computer Science)*, Vicenç Torra, Yasuo Narukawa, Beatriz López, and Mateu Villaret (Eds.), Vol. 7647. Springer, 173–184. https://doi.org/10.1007/978-3-642-34620-0_17

[43] Awanthika Senarath, Marthie Grobler, and Nalin A. G. Arachchilage. 2019. A Model for System Developers to Measure the Privacy Risk of Data. In *52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019*. University of Hawaii at Manoa, 6135–6144. https://doi.org/10.24251/HICSS.2019.738

[44] Bracha Shapira, Yuval Elovici, Adlay Meshiach, and Tsvi Kuflik. 2005. PRAW - A PRivAcy model for the Web. *Journal of the American Society for Information Science and Technology* 56, 2 (2005), 159–172. https://doi.org/10.1002/asi.20107

[45] Jitesh Shetty and J. Adibi. 2004. The Enron Email Dataset Database Schema and Brief Statistical Report. *Information sciences institute technical report, University of Southern California* 4 (2004), 120–128.

[46] Yan Shvartzshnaider, Zvonimir Pavlinovic, Ananth Balashankar, Thomas Wies, Lakshminarayanan Subramanian, Helen Nissenbaum, and Prateek Mittal. 2019. VACCINE: Using Contextual Integrity For Data Leakage Detection. In *The World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 1702–1712. https://doi.org/10.1145/3308558.3313655

[47] Yan Shvartzshnaider, Schrasing Tong, Thomas Wies, Paula Kift, Helen Nissenbaum, Lakshminarayanan Subramanian, and Prateek Mittal. 2016. Learning Privacy Expectations by Crowdsourcing Contextual Informational Norms. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA*, Arpita Ghosh and Matthew Lease (Eds.). AAAI Press, 209–218. http://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/view/14025

[48] Daniel J Solove. 2011. *Nothing to Hide: The False Tradeoff Between Privacy and Security*. Yale University Press. https://books.google.fi/books?id=UUdQi4FxRxAC

[49] Jessica Su, Ansh Shukla, Sharad Goel, and Arvind Narayanan. 2017. De-Anonymizing Web Browsing Data with Social Networks. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1261–1269. https://doi.org/10.1145/3038912.3052714

[50] Virpi Kristiina Tuunainen, Olli Pitkänen, and Marjaana Hovi. 2009. Users' Awareness of Privacy on Online Social Networking Sites - Case Facebook. In *22nd Bled eConference: eEnablement:Facilitating an Open, Effective and Representative eSociety, Bled, Slovenia, June 14-17, 2009*. Association for Information Systems, 42. http://aisel.aisnet.org/bled2009/42

[51] Xiaolei Wang, Andrea Continella, Yuexiang Yang, Yongzhong He, and Sencun Zhu. 2019. LeakDoctor: Toward Automatically Diagnosing Privacy Leaks in Mobile Applications. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3, 1 (March 2019), 28:1–28:25. https://doi.org/10.1145/3314415

[52] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I Regretted the Minute I Pressed Share": A Qualitative Study of Regrets on Facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security (SOUPS '11)*. ACM, New York, NY, USA, Article 10, 16 pages. https://doi.org/10.1145/2078827.2078841

[53] Mark Warr. 2014. *We Sacrifice Freedom for Safety, and We Need Not Do So*. UT News. https://news.utexas.edu/2014/04/21/we-sacrifice-freedom-for-safety-and-we-need-not-do-so/ Accessed: 26-10-2020.

[54] Mengwei Xu, Feng Qian, Qiaozhu Mei, Kang Huang, and Xuanzhe Liu. 2018. DeepType: On-Device Deep Learning for Input Personalization Service with Minimal Privacy Concern. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 4, Article Article 197 (Dec. 2018), 26 pages. https://doi.org/10.1145/3287075

[55] Jianhua Yin and Jianyong Wang. 2014. A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 233–242. https://doi.org/10.1145/2623330.2623715